**Review of „Weather Type Reconstruction using Machine Learning Approaches"**

**General comments:**

This study uses machine learning methods to reconstruct the CAP9 weather type classification for Europe back to the year 1728 based on station observations. Four different machine learning methods are tested (multinomial logistic regression, random forest, feedforward neural network, RNN/CNN) and compared to a reconstruction method based on Mahalanobis distance and to the original CAP9 time series published by MeteoSwiss for the reference period 1957-2020.

I find this study to be interesting, well structured and well written, and with high scientific quality of the methods and results presented.

My main objective is that the scientific relevance should be better emphasized. The authors should better explain why a weather type classification based on station observations is beneficial, especially in light of the gridded EKF400v2 reanalysis product, which goes back to the year 1602.

We'd like to thank the reviewer for this detailed review and generally positive assessment of our manuscript.

Regarding the scientific relevance, long-term WT reconstructions like the one presented in our work allows to trace decadal to multi-decadal variability and long-term changes of synoptic circulation patterns over 300 years (indicated in L. 30 f). Furthermore, with the link between WTs and surface processes, our work opens the door for a great amount of climatological analyses, e.g. for deriving a flood probability index (see Brönnimann et al., 2019) or other weather extremes.

As described in lines 32 ff in the manuscript, reanalysis datasets may be very well used for WT reconstruction. The issue, however, is the resolution of the input data which must be daily or even finer in order to allow reconstructing a classification of synoptic atmospheric patterns. EKF400v2, as well as many other reconstructions of past weather and climate going beyond the 19$^{th}$ century, only provide monthly information. For WT reconstructions going this far back in time, station observations or weather diaries are the only suitable sources of information available.

We'll try to emphasize this better in the revised manuscript, together with the scientific relevance of our work in general.

**Specific questions**

Line 47: "Whereas common statistical approaches seem to have reached their limit for this purpose, (…)". Why have they reached their limit? Please explain this better.

Thank you for this question. The limitations refer to previous WT reconstructions using common statistical approaches described in L. 38 ff. We'll rephrase this sentence in the revised manuscript: „Whereas common statistical approaches have been effective in capturing prominent atmospheric patterns, their ability to handle more complex, non-linear relationships and incorporate qualitative data is limited. Supervised machine learning (ML) classification methods offer a promising alternative, as they are well-suited for recognizing intricate non-linear patterns in atmospheric variables."

Figure 1 (Right): What is the unit of the average monthly occurrence? Days or counts?

Thank you. We'll indicate the units (number of days per month) in the y-axis label of Fig. 1 in the revised manuscript.

Table 1: Please explain what exactly the temporal pressure gradient is and how it is derived from the historical station observations.

The temporal pressure gradient is explained in L. 126 ff. We'll add a reference to Table 1 and the variable $\Delta p$ at these lines to make this clearer in the revised manuscript

Line 32ff./Line 154: I was a bit surprised to learn about the EKF400v2 reanalysis product in Line 154, which covers the period 1603-2003 and was not mentioned during the Introduction. What is the point of deriving weather type classification from station observations if a gridded reanalysis product is available for the earliest period of your observations and even before? This is in direct contradiction to the statements in line 32ff. and thus to the motivation for this paper: “With the newest generation of reanalysis datasets, many WT records could already be extended back to the 19th century (…).” and “(…) the limit for WT classifications based on atmospheric fields is set by the 20th Century Reanalysis version 3 (…), which extends back to 1806”. Please correct these statements in the Introduction and revise the motivation for a classification based on station observations in light of the available EKF400v2 reanalysis going back to 1603.

Thank you for this comment. As mentioned in our response to the reviewer's introductory comment, EKF400v2 unfortunately only provides monthly data and is thus not suitable for reconstructing daily weather types.

Line 193: Which variables are the five predictors?

Thanks for this question. The limitation to five predictors refers to a general limitation of the number of predictors to avoid multicollinearity and overfitting, rather than an a-priori choice of certain variables. The optimal combination of predictor variables is only determined during model training and described in the results section (Sect. 3.1) in L. 293 ff.

Chapter 2.3.3 and 2.3.4: What is the structure of the input layer? The Appendix says 6,8,9 x None (Time). Are these the number of stations used? What variables are used? In general, I miss a better description of the input variables of the machine learning methods. Are temperature and pressure time series used at all stations? What about the temporal pressure gradient? Please specify the structure of your input layers.

Thank you for this comment and thank you also for having a close look at the supplement. The structure of the input layer of the feedforward neural networks can be understood as a table of time series with one dimension (columns) with a length equal to the number of time series used as input (all stations and variables, e.g. 6 for the 1728 station set, see Fig. 2) and the other one with a length equal to the length of the time series (or batch). Your comment brought to light an important point that unfortunately went missing during the iterations of reworking our manuscript: whereas for the model intercomparison in Sect. 3.1 we used temporal pressure gradients as input (see L. 126 f) consistent with the baseline approach, those gradients were omitted for the NN used for WT reconstruction (Sect. 3.2 and 3.3) as tests (not shown) did not reveal consistent improvements by adding this variable. We will reinsert this statement in Sect. 2.2 and describe the input variables of the individual ML methods in Sect. 2.3 in the revised manuscript.

Chapter 2.3.3 and 2.3.4: Is the lat/lon information of the stations used as input as well? Does the machine learning model have any information on the position of the time series? If not, please discuss this.

Thank you for this question. We did not include lat/lon information or other direct information on the station location as input to the model. As supervised machine learning methods are designed to identify patterns in the input data (i.e. station observations) related to a given category (in this case with given circulation patterns (WTs) over a certain region), we expect the models to find the relevant spatial patterns even without the knowledge of the exact position of the stations. Tests were made with indirect spatial information, using spatial gradients between stations (e.g. the pressure difference between Stockholm and Milan for a north-south gradient). However, they did not show any benefits with the tested models (not shown) supporting the former statement. From the presented methods, CNNs would be the most appropriate to include the spatial dimension directly, treating the station observations as cells of a spatial grid. However, this would need further research related e.g. to the grid structure and imputation of missing cells. As the validation metrics showed good results without spatial information, we did not pursue this issue in our study.

Line 241f. What are "(…) all available pressure and temperature series"? Please specify.

Thanks for this question. As stated in L. 240 f, we tested subsets of stations or variables (e.g. only pressure / only temperature, subset of stations to achieve a more equal spatial distribution). „All" in this case means that we use the full set of stations and variables available. We will try to make this clearer in the revised manuscript.

Line 278: What is the advantage of the Heidke skill score? How can it be interpreted compared to overall accuracy?

The Heidke skill score includes an important aspect that the overall accuracy alone does not indicate: as described in L. 280 ff, the HSS is calculated for each WT individually and thus accounts for differences in the occurrence frequency of the individual WTs. Overall accuracy, however, may weight more frequent WTs stronger. Therefore, as an example, a high accuracy together with lower HSS values allows for the interpretation that prediction errors might originate from individual WTs. Furthermore, the HSS provides a value with respect to a reference (forecast by chance).

Line 353: "(…) which are mostly within the range of uncertainty of model training." How do you quantify the range of uncertainty of model training to reach this conclusion?

Thanks for this remark. As a rough measure for the uncertainty of model training we took the variance of the validation metrics of the dataset splits (outer folds). The variance of accuracy and HSS in the outer folds were larger than the improvements gained by adding wet days as additional predictors. We'll indicate this in the revised manuscript.

Line 368f.: "The accuracy for the earliest period between 01.01.1728 and 31.12.1737 is already remarkably high with a value of 77.8 % despite the limited set of available stations." This sentence is misleading, because it suggests that you know the accuracy of your model for the earliest period. But you can't estimate the accuracy of the early period, because you don't have labels for that time to which you could compare your classifications to. If I got it right, the 77.8% indicate the accuracy of your trained model for a test set from the period 1957-2020 compared to the MeteoSwiss time series, whereby your model uses the number of stations only that are available since 1728. But your actual accuracy in the early period could be lower than that due to lower data quality in the early period e.g. measurement errors. Please refine the statement and discuss the data quality within your time series.

This is an excellent comment. We agree that the phrasing of this sentence may be misleading. We will change this in the revised manuscript and add a sentence on the accuracy in the reference period vs. the actual accuracy in the past at L. 370: „The achieved accuracy using the smallest station set (stations available from 01.01.1728 to 31.12.1737) is already remarkably high […]. Adding more stations […]. Note that validation metrics shown in Table 3 only provide values with respect to the reference period 1957-2020. The actual values for the past periods may be lower due to larger uncertainties and errors in the data, but unfortunately cannot be determined due to the lack of a historical reference WT series.“

Figure 5: The plots are quite small and hard to compare by eye. It could help to increase the size and/or to show the differences of the false composites and the true composites to obs composites in order to better show the differences in the pressure fields. I'm also wondering how many cases each composite plot is derived from. The numbers could be indicated above the plots.

Thank you for these suggestions. We will increase the size of the plot and indicate the number of cases in a revised figure. We originally also considered showing deviation maps of true / false composites with respect to observations (see also response to RC3). However, we found this to blur information on the position of low and high pressure systems in the true / false prediction maps. Although less apparent than when showing deviations, the discussed differences between the maps can still be determined from the absolute values shown in Fig. 5, thus we deemed this solution to be better.

Discussion: I miss a discussion on why including the previous days in the RNN/CNN setup didn't help to improve the accuracy of the weather type classification. Is this in line with what the authors expected? What could be the reasons for this?

Thanks for this remark. Whereas we expected some improvement when taking the previous days into consideration, the available data for a certain day including the pressure gradient with respect to the day before (as used in NN) seems to be sufficient for correctly determining the corresponding WT. Added value of temporal series may be linked to information on preferential transitions which can complement station observations (arguably the case for few station series, see Table 2). We will add such a statement in L. 331 in the revised manuscript. The main sources of error (i.e. the reason for 10-20% wrongly assigned WTs), however, seem to have a different origin and cannot be solved by using data from previous days (e.g. the spatial coverage by stations, see L. 312 f, L. 480).

Supplement Table S2.1: Please explain the variable names

Thanks for this suggestion. We will explain the variable names in the table caption in the revised supplement.